



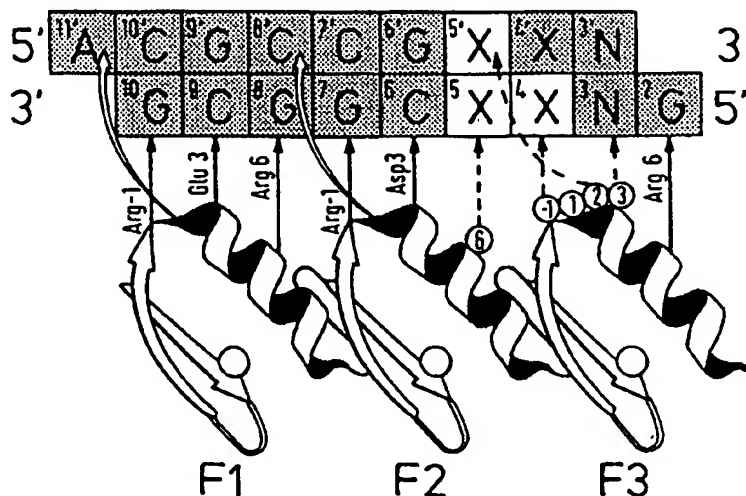
## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup>:</b> <b>C12N 15/10, 15/12, C07K 1/04, C12N 15/62, C12Q 1/68, C07K 14/47, A61K 48/00</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 98/53057</b> <b>(43) International Publication Date:</b> 26 November 1998 (26.11.98)
<b>(21) International Application Number:</b> PCT/GB98/01510 <b>(22) International Filing Date:</b> 26 May 1998 (26.05.98) <b>(30) Priority Data:</b> 9710809.6 23 May 1997 (23.05.97) GB <b>(71) Applicant (for all designated States except US):</b> MEDICAL RESEARCH COUNCIL [GB/GB]; 20 Park Crescent, London W1N 4AL (GB). <b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only):</b> CHOO, Yen [GR/GB]; MRC Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH (GB). KLUG, Aaron [GB/GB]; MRC Laboratory of Molecular Biology, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH (GB). ISALAN, Mark [GB/GB]; 24 Shottfield Avenue, East Sheen, London SW14 8EA (GB). <b>(74) Agents:</b> MASCHIO, Antonio et al.; D. Young & Co., 21 New Fetter Lane, London EC4A 1DA (GB).	<b>(81) Designated States:</b> AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, GW, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).  <b>Published</b> <i>With international search report.</i> <i>Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>	

(54) Title: NUCLEIC ACID BINDING POLYPEPTIDE LIBRARY

## (57) Abstract

The invention relates to a zinc finger polypeptide library in which each polypeptide comprises more than one zinc finger which has been at least partially randomised, and to a set of zinc finger polypeptide libraries which encode overlapping zinc finger polypeptides, each polypeptide comprising more than one zinc finger which has been at least partially randomised, and which polypeptides may be assembled after selection to form a multifinger zinc finger polypeptide.



**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BR	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

## NUCLEIC ACID BINDING POLYPEPTIDE LIBRARY

The present invention relates to a library system for the selection of zinc finger polypeptides. In particular, the invention relates to a binary system, in which zinc  
5 finger motifs are randomised in overlapping regions and to smart libraries incorporating limited directed randomisation at selected positions.

Protein-nucleic acid recognition is a commonplace phenomenon which is central to a large number of biomolecular control mechanisms which regulate the functioning of  
10 eukaryotic and prokaryotic cells. For instance, protein-DNA interactions form the basis of the regulation of gene expression and are thus one of the subjects most widely studied by molecular biologists.

A wealth of biochemical and structural information explains the details of protein-DNA  
15 recognition in numerous instances, to the extent that general principles of recognition have emerged. Many DNA-binding proteins contain independently folded domains for the recognition of DNA, and these domains in turn belong to a large number of structural families, such as the leucine zipper, the "helix-turn-helix" and zinc finger families.

20

Despite the great variety of structural domains, the specificity of the interactions observed to date between protein and DNA most often derives from the complementarity of the surfaces of a protein  $\alpha$ -helix and the major groove of DNA [Klug, (1993) Gene 135:83-92]. In light of the recurring physical interaction of  $\alpha$ -helix  
25 and major groove, the tantalising possibility arises that the contacts between particular amino acids and DNA bases could be described by a simple set of rules; in effect a stereochemical recognition code which relates protein primary structure to binding-site sequence preference.

30 It is clear, however, that no code will be found which can describe DNA recognition by all DNA-binding proteins. The structures of numerous complexes show significant

differences in the way that the recognition  $\alpha$ -helices of DNA-binding proteins from different structural families interact with the major groove of DNA, thus precluding similarities in patterns of recognition. The majority of known DNA-binding motifs are not particularly versatile, and any codes which might emerge would likely describe  
5 binding to a very few related DNA sequences.

Even within each family of DNA-binding proteins, moreover, it has hitherto appeared that the deciphering of a code would be elusive. Due to the complexity of the protein-DNA interaction, there does not appear to be a simple "alphabetic" equivalence  
10 between the primary structures of protein and nucleic acid which specifies a direct amino acid to base relationship.

International patent application WO 96/06166 addresses this issue and provides a "syllabic" code which explains protein-DNA interactions for zinc finger nucleic acid  
15 binding proteins. A syllabic code is a code which relies on more than one feature of the binding protein to specify binding to a particular base, the features being combinable in the forms of "syllables", or complex instructions, to define each specific contact.

20 However, this code is incomplete, providing no specific instructions permitting the specific selection of nucleotides other than G in the 5' position of each triplet. The method relies on randomisation and subsequent selection in order to generate nucleic acid binding proteins for other specificities. Even with the aid of partial randomisation and selection, however, neither the method reported in WO 96/06166 nor any other  
25 methods of the prior art have succeeded in isolating a zinc finger polypeptide based on the first finger of Zif268 capable of binding triplets wherein the 5' base is other than G or T. This is a serious shortfall in any ability to design zinc finger proteins.

Moreover, this document relies upon the notion that zinc fingers bind to a nucleic acid  
30 triplet or multiples thereof, as does all of the prior art. We have now determined that zinc finger binding sites are determined by overlapping 4 bp subsites, and that

sequence-specificity at the boundary between subsites arises from synergy between adjacent fingers. This has important implications for the design and selection of zinc fingers with novel DNA binding specificities.

## 5 Summary of the Invention

The present invention recognises the importance of overlapping 4 bp subsite recognition in zinc finger polypeptide design. The resultant synergy between zinc fingers is overlooked in classical zinc finger library design, in which only a single zinc finger is  
10 randomised in each library.

Accordingly, the present invention provides a zinc finger polypeptide library in which each polypeptide comprises more than one zinc finger which has been at least partially randomised.

15 Preferably, the invention provides a group of zinc finger polypeptide libraries which encode overlapping zinc finger polypeptides, each polypeptide comprising more than one zinc finger which has been at least partially randomised, and which polypeptides may be assembled after selection to form a multifinger zinc finger polypeptide.

20 In a further aspect, the invention relates to a library as described above in which randomisation is limited to substituting amino acids which are known to dictate variation in binding site specificity. The present invention provides a code of amino acid position bias which permits the selection of the library against any nucleic acid  
25 sequence as the target sequence, and the production of a specific nucleic acid-binding protein which will bind thereto. Moreover, the invention provides a method by which a zinc finger protein specific for any given nucleic acid sequence may be designed and optimised. The present invention therefore concerns a recognition bias which has been elucidated for the interactions of classical zinc fingers with nucleic acid. In this case a  
30 pattern of rules is provided which covers binding to all nucleic acid sequences.

The code set forth in the present invention takes account of synergistic interactions between adjacent zinc fingers, thereby allowing the selection of any desired binding site.

5

### **Brief Description of the Drawings**

Figure 1 illustrates zinc finger-DNA interactions. A: model of classical triplet interactions with DNA base triplets in Zif268; B: similar model showing quadruplet  
10 interactions; C: model of library design for recognition code determination.

Figure 2 shows the amino acid sequence of three fingers used for phage display selection in the determination of recognition code.

15 Figure 3 lists the sequence-specific zinc finger clones obtained from phage selections, and their binding site signatures.

Figure 4 shows the base/amino acid correlation of the clones isolated from phage selections. Recognition patterns are highlighted.

20

Figure 5 illustrates the sequence-specific interactions selected for at position 2 of the  $\alpha$ -helix, binding to position 1 of the quadruplet.

25 Figure 6 is a schematic diagram of the construction of a library according to the invention.

**Detailed Description of the Invention**

The present invention relates to libraries. The term "library" is used according to its common usage in the art, to denote a collection of polypeptides or, preferably, nucleic acids encoding polypeptides. The polypeptides of the invention contain regions of randomisation, such that each library will comprise or encode a repertoire of polypeptides, wherein individual polypeptides differ in sequence from each other. The same principle is present in virtually all libraries developed for selection, such as by phage display.

10

Randomisation, as used herein, refers to the variation of the sequence of the polypeptides which comprise the library, such that various amino acids may be present at any given position in different polypeptides. Randomisation may be complete, such that any amino acid may be present at a given position, or partial, such that only certain amino acids are present. Preferably, the randomisation is achieved by mutagenesis at the nucleic acid level, for example by synthesising novel genes encoding mutant proteins and expressing these to obtain a variety of different proteins. Alternatively, existing genes can be themselves mutated, such by site-directed or random mutagenesis, in order to obtain the desired mutant genes.

20

Mutations may be performed by any method known to those of skill in the art. Preferred, however, is site-directed mutagenesis of a nucleic acid sequence encoding the protein of interest. A number of methods for site-directed mutagenesis are known in the art, from methods employing single-stranded phage such as M13 to PCR-based techniques (see "PCR Protocols: A guide to methods and applications", M.A. Innis, D.H. Gelfand, J.J. Sninsky, T.J. White (eds.). Academic Press, New York, 1990). Preferably, the commercially available Altered Site II Mutagenesis System (Promega) may be employed, according to the directions given by the manufacturer.

30 Screening of the proteins produced by mutant genes is preferably performed by expressing the genes and assaying the binding ability of the protein product. A simple

and advantageously rapid method by which this may be accomplished is by phage display, in which the mutant polypeptides are expressed as fusion proteins with the coat proteins of filamentous bacteriophage, such as the minor coat protein pII of bacteriophage m13 or gene III of bacteriophage Fd, and displayed on the capsid of bacteriophage transformed with the mutant genes. The target nucleic acid sequence is used as a probe to bind directly to the protein on the phage surface and select the phage possessing advantageous mutants, by affinity purification. The phage are then amplified by passage through a bacterial host, and subjected to further rounds of selection and amplification in order to enrich the mutant pool for the desired phage and eventually isolate the preferred clone(s). Detailed methodology for phage display is known in the art and set forth, for example, in US Patent 5,223,409; Choo and Klug, (1995) Current Opinions in Biotechnology 6:431-436; Smith, (1985) Science 228:1315-1317; and McCafferty *et al.*, (1990) Nature 348:552-554; all incorporated herein by reference. Vector systems and kits for phage display are available commercially, for example from Pharmacia.

The polypeptides which comprise the libraries according to the invention are zinc finger polypeptides. In other words, they comprise a Cys2-His2 zinc finger motif. It is a feature of the invention that each polypeptide comprises more than one zinc finger, such that the library may be selected on the basis of the interaction between two or more zinc fingers on the polypeptide.

Zinc fingers, as is known in the art, are nucleic acid binding molecules. Each zinc finger binds to a quadruplet sequence in a target nucleic acid through contacts between specific amino acid residues of the  $\alpha$ -helix of the zinc finger and the nucleic acid strand. The quadruplets specified in the present invention are overlapping, such that, when read 3' to 5' on the -strand of the nucleic acid, base 4 of the first quadruplet is base 1 of the second, and so on. Accordingly, in the present application, the bases of each quadruplet are referred by number, from 1 to 4, 1 being the 3' base and 4 being the 5' base. Base 4 is equivalent to the 5' base of a classical zinc finger binding triplet. In general, base 4 is bound through a contact at position +6 of the  $\alpha$ -helix, base 3



through a contact at position +3, base 2 through a contact at position -1 and base 1 through a contact to the opposite strand of double-stranded nucleic acids at position +2.

- 5 All of the nucleic acid-binding residue positions of zinc fingers, as referred to herein, are numbered from the first residue in the  $\alpha$ -helix of the finger, ranging from +1 to +9. "-1" refers to the residue in the framework structure immediately preceding the  $\alpha$ -helix in a Cys2-His2 zinc finger polypeptide.
- 10 Residues referred to as "+ +2" are residues present in an adjacent (C-terminal) finger. They reflect the synergistic cooperation between position +2 on base 1 (on the + strand) and position +6 of the preceding (N-terminal) finger on base 4 of the preceding (3') quadruplet, which is the same base due to the overlap. Where there is no C-terminal adjacent finger, "+ +" interactions do not operate.
- 15 Cys2-His2 zinc finger binding proteins, as is well known in the art, bind to target nucleic acid sequences via  $\alpha$ -helical zinc metal atom co-ordinated binding motifs known as zinc fingers. Each zinc finger in a zinc finger nucleic acid binding protein is responsible for determining binding to a nucleic acid quadruplet in a nucleic acid
- 20 binding sequence. Preferably, there are 2 or more zinc fingers, for example 2, 3, 4, 5 or 6 zinc fingers, in each binding protein. Advantageously, there are 3 zinc fingers in each zinc finger binding protein.

- The present invention allows the production of what are essentially artificial nucleic
- 25 acid binding proteins. In these proteins, artificial analogues of amino acids may be used, to impart the proteins with desired properties or for other reasons. Thus, the term "amino acid", particularly in the context where "any amino acid" is referred to, means any sort of natural or artificial amino acid or amino acid analogue that may be employed in protein construction according to methods known in the art. Moreover,
- 30 any specific amino acid referred to herein may be replaced by a functional analogue thereof, particularly an artificial functional analogue. The nomenclature used herein

therefore specifically comprises within its scope functional analogues of the defined amino acids.

The  $\alpha$ -helix of a zinc finger binding protein aligns antiparallel to the nucleic acid strand, such that the primary nucleic acid sequence is arranged 3' to 5' in order to correspond with the N terminal to C-terminal sequence of the zinc finger. Since nucleic acid sequences are conventionally written 5' to 3', and amino acid sequences N-terminus to C-terminus, the result is that when a nucleic acid sequence and a zinc finger protein are aligned according to convention, the primary interaction of the zinc finger is with the - strand of the nucleic acid, since it is this strand which is aligned 3' to 5'. These conventions are followed in the nomenclature used herein. It should be noted, however, that in nature certain fingers, such as finger 4 of the protein GLI, bind to the + strand of nucleic acid: see Suzuki *et al.*, (1994) NAR 22:3397-3405 and Pavletich and Pabo, (1993) Science 261:1701-1707. The incorporation of such fingers into nucleic acid binding molecules according to the invention is envisaged.

The libraries of the present invention allow selection for synergistic cooperation between adjacent zinc fingers by promoting coselection of adjacent fingers against a single DNA target. This is achieved by randomising, in the same zinc finger polypeptide, more than one zinc finger. In a preferred embodiment, approximately one and a half zinc fingers are randomised in each polypeptide, but this may be varied according to library design.

The zinc finger polypeptides encoded in the library of the invention may comprise any number of zinc fingers, provided this is more than one. Advantageously, each polypeptide encodes between three and six zinc fingers. In each library, the randomisation extends to cover the overlap of at least one pair of zinc fingers. Preferably, the overlap of a single pair is covered.

Preferably, the libraries of the present invention are provided as sets. Thus, a three zinc finger polypeptide comprising fingers F1, F2 and F3 may be presented in a set of

two libraries, each library comprising a two zinc finger polypeptide. A first library is composed of polypeptides consisting essentially of F1 and F2, whilst a second library is composed of polypeptides consisting essentially of F2 and F3. The randomisation in each library includes the overlap between F1 and F2, and F2 and F3 respectively.

5

Preferably, each library will comprise randomisation at at least position 6 of a first finger and position 2 of a second finger. Since these residues contact the same base pair on a double stranded nucleic acid target, it is advantageous that they be varied together.

10

In the case of a three zinc finger polypeptide, the first library will be randomised in fingers F1 and F2, whilst the second is randomised in F2 and F3. Polypeptides may be recombined, post-selection, in the F2 sequence to create a single polypeptide containing F1, F2 and F3. This polypeptide will have been selected taking into account

15

Advantageously, a greater number of position may be varied in each zinc finger. Preferably, residues selected from positions -1, 1, 2, 3 5 and 6 are varied in a first zinc finger and positions -1, 1, 2 and 3 in a second. In a companion library, positions 3, 5 and 6 may be varied in the second finger, and positions -1, 1, 2 and 3 in a third finger. In the final finger (in the case of a three finger protein this will be the third finger), residues 5 and 6 may also be varied.

20

In order that the libraries may be recombined after selection, the polypeptides are preferably designed to include a suitable restriction site in the nucleic acid encoding the zinc finger shared by two libraries. The position of the cleavage site will dictate the precise site of the variations made in the shared zinc finger in each library. Thus, in a set of two libraries encoding a three zinc finger protein, if the cleavage site is between positions 3 and 5 of the  $\alpha$ -helix, positions 3 and 5 may be randomised in a first library and positions 5 and 6 in a second.

25  
30

Although it is preferred that residues for randomisation or variation be selected from positions -1, 1, 2, 3, 5 and 6, further residues may also be randomised. For example, the randomisation of position 8 may be advantageous. Moreover, it is envisaged that fewer than all of the given positions are randomised.

5

In a preferred embodiment, a two-library system for selection of a three-finger protein is varied at F1 positions -1, 2, 3, 5, and 6 and F2 positions -1, 1, 2 and 3 in the first library. The second library is varied at F2 positions 3 and 6 and F3 positions -1, 1, 2, 3, 5 and 6. In this case, the cleavage and recombination point will be between residues

10 3 and 5, preferably between residues 4 and 5, of the  $\alpha$ -helix of F2.

Subsequent to the recombination event, recombined polypeptide-encoding nucleic acids may be expressed in suitable expression systems, or cloned into Fd phage for further selection.

15

In a preferred aspect of the present invention, the libraries of the invention are not truly randomised at the selected positions, but only partially randomised so that certain but not all amino acids are encoded. This strategy may be used for two purposes.

20 In a first embodiment, variation is restricted to those amino acids which are known to be capable of directing sequence-specific binding of nucleic acid target sequences when incorporated at a given position in the  $\alpha$ -helix of a zinc finger. It is known that certain amino acids are not suitable for incorporation at certain positions, irrespective of target sequence. These amino acids are avoided.

25

In a second embodiment, variation is restricted to those amino acids which are known to be capable of directing sequence-specific binding of nucleic acid target sequences when incorporated at a given position in the  $\alpha$ -helix of a zinc finger, and variation is directed to specify those residues which are known to favour binding to a specific target

30 sequence at any given position. Thus, the invention permits the design of dedicated

libraries from which polypeptides capable of binding to specific target sequence, or to a series of related target sequences, may be selected.

In the first embodiment, which provides a library system for general application, randomisation is preferably effected at all of the positions indicated above. Preferably, the amino acids selected to appear at each given position are as set forth in Table 1:

Position	Possible Amino Acids
-1	R, Q, H, N, D, A, T
1	S, R, K, N
2	D, A, R, Q, H, K, S, N
3	H, N, S, T, V, A, D
5	I, T, K
6	R, Q, V, A, E, K, N, T

**TABLE 1**

10

It is not necessary for each finger to be randomised at each of the positions given in table 1. In a preferred embodiment, a library for selecting a three-finger protein is constructed according to the specifications given in Table 2:

15

Library 1		Library 2	
<b>F1:</b>	<b>amino acid</b>	<b>F1:</b>	<b>amino acid</b>
-1	R, Q, H, N, D, A		
2	D, A, R, Q, H, K, S, N		
3	H, N, S, T, V, A, D		
5	I, T		
6	R, Q, V, A, E, K, N, T		
<b>F2</b>			
-1	R, Q, H, N, D, A, T		
1	S, R		
2	D, A, R, Q, H, K, S, N		
3	H, N, S, T, V, A, D	3	H, N, S, T, V, A, D
		6	R, Q, V, A, E, K, N, T
<b>F3</b>			
		-1	R, Q, H, N, D, A, T
		1	R, K, S, N
		2	D, A, R, Q, H, K, S, N
		3	H, N, S, T, V, A, D
		5	K, I, T
		6	R, Q, V, A, E, K, N, T

TABLE 2

In the second embodiment, the identity of each amino acid at any particular position is selected according to zinc finger recognition rules as provided herein. In a preferred aspect, therefore, the invention provides a method for preparing a nucleic acid binding protein of the Cys2-His2 zinc finger class capable of binding to a nucleic acid quadruplet in a target nucleic acid sequence, wherein binding to each base of the

quadruplet by an  $\alpha$ -helical zinc finger nucleic acid binding motif in the protein is determined as follows:

- a) if base 4 in the quadruplet is G, then position +6 in the  $\alpha$ -helix is Arg or Lys;
- 5 b) if base 4 in the quadruplet is A, then position +6 in the  $\alpha$ -helix is Glu, Asn or Val;
- c) if base 4 in the quadruplet is T, then position +6 in the  $\alpha$ -helix is Ser, Thr, Val or Lys;
- d) if base 4 in the quadruplet is C, then position +6 in the  $\alpha$ -helix is Ser, Thr, Val, Ala, Glu or Asn;
- 10 e) if base 3 in the quadruplet is G, then position +3 in the  $\alpha$ -helix is His;
- f) if base 3 in the quadruplet is A, then position +3 in the  $\alpha$ -helix is Asn;
- g) if base 3 in the quadruplet is T, then position +3 in the  $\alpha$ -helix is Ala, Ser or Val; provided that if it is Ala, then one of the residues at -1 or +6 is a small residue;
- h) if base 3 in the quadruplet is C, then position +3 in the  $\alpha$ -helix is Ser, Asp, Glu,
- 15 Leu, Thr or Val;
- i) if base 2 in the quadruplet is G, then position -1 in the  $\alpha$ -helix is Arg;
- j) if base 2 in the quadruplet is A, then position -1 in the  $\alpha$ -helix is Gln;
- k) if base 2 in the quadruplet is T, then position -1 in the  $\alpha$ -helix is His or Thr;
- l) if base 2 in the quadruplet is C, then position -1 in the  $\alpha$ -helix is Asp or His.
- 20 m) if base 1 in the quadruplet is G, then position +2 is Glu;
- n) if base 1 in the quadruplet is A, then position +2 Arg or Gln;
- o) if base 1 in the quadruplet is C, then position +2 is Asn, Gln, Arg, His or Lys;
- p) if base 1 in the quadruplet is T, then position +2 is Ser or Thr.
- 25 The foregoing represents a set of rules which permits the design of a zinc finger binding protein specific for any given nucleic acid sequence. A novel finding related thereto is that position +2 in the helix is responsible for determining the binding to base 1 of the quadruplet. In doing so, it cooperates synergistically with position +6, which determines binding at base 4 in the quadruplet, bases 1 and 4 being overlapping
- 30 in adjacent quadruplets.

Although zinc finger polypeptides are considered to bind to overlapping quadruplet sequences, the method of the present invention allows polypeptides to be designed to bind to target sequences which are not multiples of overlapping quadruplets. For example, a zinc finger polypeptide may be designed to bind to a palindromic target sequence. Such sequences are commonly found as, for example, restriction enzyme target sequences.

Preferably, creation of zinc fingers which bind to fewer than three nucleotides is achieved by specifying, in the zinc finger, amino acids which are unable to support H-bonding with the nucleic acid in the relevant position.

Advantageously, this is achieved by substituting Gly at position -1 (to eliminate a contact with base 2) and/or Ala at positions +3 and/or +6 (to eliminate contacts at the 3rd or 4th base respectively).

Preferably, the contact with the final (3') base in the target sequence should be strengthened, if necessary, by substituting a residue at the relevant position which is capable of making a direct contact with the phosphate backbone of the nucleic acid.

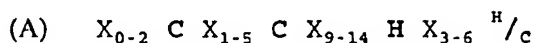
These and other considerations may be incorporated in a library set in accordance with the invention.

A zinc finger binding motif is a structure well known to those in the art and defined in, for example, Miller *et al.*, (1985) EMBO J. 4:1609-1614; Berg (1988) PNAS (USA) 85:99-102; Lee *et al.*, (1989) Science 245:635-637; see International patent applications WO 96/06166 and WO 96/32475, corresponding to USSN 08/422,107, incorporated herein by reference.

As used herein, "nucleic acid" refers to both RNA and DNA, constructed from natural nucleic acid bases or synthetic bases, or mixtures thereof. Preferably, however, the binding proteins of the invention are DNA binding proteins.



In general, a preferred zinc finger framework has the structure:

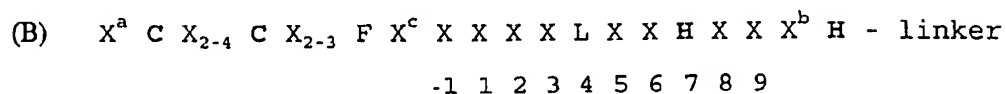


5

where X is any amino acid, and the numbers in subscript indicate the possible numbers of residues represented by X.

In a preferred aspect of the present invention, zinc finger nucleic acid binding motifs may be represented as motifs having the following primary structure:

10



15

wherein X (including  $X^a$ ,  $X^b$  and  $X^c$ ) is any amino acid.  $X_{2-4}$  and  $X_{2-3}$  refer to the presence of 2 or 4, or 2 or 3, amino acids, respectively. The Cys and His residues, which together co-ordinate the zinc metal atom, are marked in bold text and are usually invariant, as is the Leu residue at position +4 in the  $\alpha$ -helix.

20

Modifications to this representation may occur or be effected without necessarily abolishing zinc finger function, by insertion, mutation or deletion of amino acids. For example it is known that the second His residue may be replaced by Cys (Krizek *et al.*, (1991) J. Am. Chem. Soc. 113:4518-4523) and that Leu at +4 can in some circumstances be replaced with Arg. The Phe residue before  $X_c$  may be replaced by any aromatic other than Trp. Moreover, experiments have shown that departure from the preferred structure and residue assignments for the zinc finger are tolerated and may even prove beneficial in binding to certain nucleic acid sequences. Even taking this into account, however, the general structure involving an  $\alpha$ -helix co-ordinated by a zinc atom which contacts four Cys or His residues, does not alter. As used herein, structures (A) and (B) above are taken as an exemplary structure representing all zinc

25

30

finger structures of the Cys2-His2 type.

Preferably,  $X^a$  is  $F/Y-X$  or  $P-F/Y-X$ . In this context, X is any amino acid. Preferably, in this context X is E, K, T or S. Less preferred but also envisaged are Q, V, A and P. The remaining amino acids remain possible.

5

Preferably,  $X_{2-4}$  consists of two amino acids rather than four. The first of these amino acids may be any amino acid, but S, E, K, T, P and R are preferred. Advantageously, it is P or R. The second of these amino acids is preferably E, although any amino acid may be used.

10

Preferably,  $X^b$  is T or I.

Preferably,  $X^c$  is S or T.

- 15 Preferably,  $X_{2-3}$  is G-K-A, G-K-C, G-K-S or G-K-G. However, departures from the preferred residues are possible, for example in the form of M-R-N or M-R.

Preferably, the linker is T-G-E-K or T-G-E-K-P.

- 20 As set out above, the major binding interactions occur with amino acids -1, +2, +3 and +6. Amino acids +4 and +7 are largely invariant. The remaining amino acids may be essentially any amino acids. Preferably, position +9 is occupied by Arg or Lys. Advantageously, positions +1, +5 and +8 are not hydrophobic amino acids, that is to say are not Phe, Trp or Tyr.

25

In a most preferred aspect, therefore, bringing together the above, the invention allows the definition of every residue in a zinc finger nucleic acid binding motif which will bind specifically to a given nucleic acid quadruplet.

- 30 The code provided by the present invention is not entirely rigid; certain choices are provided. For example, positions +1, +5 and +8 may have any amino acid

allocation, whilst other positions may have certain options: for example, the present rules provide that, for binding to a central T residue, any one of Ala, Ser or Val may be used at +3. In its broadest sense, therefore, the present invention provides a very large number of proteins which are capable of binding to every defined target nucleic acid quadruplet.

Preferably, however, the number of possibilities may be significantly reduced. For example, the non-critical residues +1, +5 and +8 may be occupied by the residues Lys, Thr and Gln respectively as a default option. In the case of the other choices, for example, the first-given option may be employed as a default. Thus, the code according to the present invention allows the design of a single, defined polypeptide (a "default" polypeptide) which will bind to its target quadruplet.

In a further aspect of the present invention, there is provided a method for preparing a nucleic acid binding protein of the Cys2-His2 zinc finger class capable of binding to a target nucleic acid sequence, comprising the steps of:

- a) selecting a model zinc finger domain from the group consisting of naturally occurring zinc fingers and consensus zinc fingers; and
- b) mutating one or more of positions -1, +2, +3 and +6 of the finger as required according to the rules set forth above.

In general, naturally occurring zinc fingers may be selected from those fingers for which the nucleic acid binding specificity is known. For example, these may be the fingers for which a crystal structure has been resolved: namely Zif 268 (Elrod-Erickson *et al.*, (1996) Structure 4:1171-1180), GLI (Pavletich and Pabo, (1993) Science 261:1701-1707), Tramtrack (Fairall *et al.*, (1993) Nature 366:483-487) and YY1 (Houbaviy *et al.*, (1996) PNAS (USA) 93:13577-13582).

The naturally occurring zinc finger 2 in Zif 268 makes an excellent starting point from which to engineer a zinc finger and is preferred.

Consensus zinc finger structures may be prepared by comparing the sequences of  
5 known zinc fingers, irrespective of whether their binding domain is known. Preferably, the consensus structure is selected from the group consisting of the consensus structure P Y K C P E C G K S F S Q K S D L V K H Q R T H T G, and the consensus structure P Y K C S E C G K A F S Q K S N L T R H Q R I H T G E K P.

10 The consensuses are derived from the consensus provided by Krizek *et al.*, (1991) J. Am. Chem. Soc. 113:4518-4523 and from Jacobs, (1993) PhD thesis, University of Cambridge, UK. In both cases, the linker sequences described above for joining two zinc finger motifs together, namely TGEK or TGEKP can be formed on the ends of the consensus. Thus, a P may be removed where necessary, or, in the case of the  
15 consensus terminating T G, E K (P) can be added.

When the nucleic acid specificity of the model finger selected is known, the mutation of the finger in order to modify its specificity to bind to the target nucleic acid may be directed to residues known to affect binding to bases at which the natural and desired  
20 targets differ. Otherwise, mutation of the model fingers should be concentrated upon residues -1, +2, +3 and +6 as provided for in the foregoing rules.

In order to produce a binding protein having improved binding, moreover, the rules provided by the present invention may be supplemented by physical or virtual  
25 modelling of the protein/nucleic acid interface in order to assist in residue selection.

Zinc finger binding motifs designed according to the invention may be combined into nucleic acid binding proteins having a multiplicity of zinc fingers. Preferably, the proteins have at least two zinc fingers. In nature, zinc finger binding proteins  
30 commonly have at least three zinc fingers, although two-zinc finger proteins such as Tramtrack are known. The presence of at least three zinc fingers is preferred. Binding

proteins may be constructed by joining the required fingers end to end, N-terminus to C-terminus. Preferably, this is effected by joining together the relevant nucleic acid coding sequences encoding the zinc fingers to produce a composite coding sequence encoding the entire binding protein. The invention therefore provides a method for  
5 producing a nucleic acid binding protein as defined above, wherein the nucleic acid binding protein is constructed by recombinant DNA technology, the method comprising the steps of:

- 10 a) preparing a nucleic acid coding sequence encoding two or more zinc finger binding motifs as defined above, placed N-terminus to C-terminus;
  - b) inserting the nucleic acid sequence into a suitable expression vector; and
  - c) expressing the nucleic acid sequence in a host organism in order to obtain the nucleic acid binding protein.
- 15 A "leader" peptide may be added to the N-terminal finger. Preferably, the leader peptide is MAEEKP.

The nucleic acid encoding the nucleic acid binding protein according to the invention can be incorporated into vectors for further manipulation. As used herein, vector (or  
20 plasmid) refers to discrete elements that are used to introduce heterologous nucleic acid into cells for either expression or replication thereof. Selection and use of such vehicles are well within the skill of the person of ordinary skill in the art. Many vectors are available, and selection of appropriate vector will depend on the intended use of the vector, i.e. whether it is to be used for DNA amplification or for nucleic acid  
25 expression, the size of the DNA to be inserted into the vector, and the host cell to be transformed with the vector. Each vector contains various components depending on its function (amplification of DNA or expression of DNA) and the host cell for which it is compatible. The vector components generally include, but are not limited to, one or more of the following: an origin of replication, one or more marker genes, an enhancer  
30 element, a promoter, a transcription termination sequence and a signal sequence.

Both expression and cloning vectors generally contain nucleic acid sequence that enable the vector to replicate in one or more selected host cells. Typically in cloning vectors, this sequence is one that enables the vector to replicate independently of the host chromosomal DNA, and includes origins of replication or autonomously replicating sequences. Such sequences are well known for a variety of bacteria, yeast and viruses. The origin of replication from the plasmid pBR322 is suitable for most Gram-negative bacteria, the 2 $\mu$  plasmid origin is suitable for yeast, and various viral origins (e.g. SV 40, polyoma, adenovirus) are useful for cloning vectors in mammalian cells. Generally, the origin of replication component is not needed for mammalian expression vectors unless these are used in mammalian cells competent for high level DNA replication, such as COS cells.

Most expression vectors are shuttle vectors, i.e. they are capable of replication in at least one class of organisms but can be transfected into another class of organisms for expression. For example, a vector is cloned in *E. coli* and then the same vector is transfected into yeast or mammalian cells even though it is not capable of replicating independently of the host cell chromosome. DNA may also be replicated by insertion into the host genome. However, the recovery of genomic DNA encoding the nucleic acid binding protein is more complex than that of exogenously replicated vector because restriction enzyme digestion is required to excise nucleic acid binding protein DNA. DNA can be amplified by PCR and be directly transfected into the host cells without any replication component.

Advantageously, an expression and cloning vector may contain a selection gene also referred to as selectable marker. This gene encodes a protein necessary for the survival or growth of transformed host cells grown in a selective culture medium. Host cells not transformed with the vector containing the selection gene will not survive in the culture medium. Typical selection genes encode proteins that confer resistance to antibiotics and other toxins, e.g. ampicillin, neomycin, methotrexate or tetracycline, complement auxotrophic deficiencies, or supply critical nutrients not available from complex media.

As to a selective gene marker appropriate for yeast, any marker gene can be used which facilitates the selection for transformants due to the phenotypic expression of the marker gene. Suitable markers for yeast are, for example, those conferring resistance to antibiotics G418, hygromycin or bleomycin, or provide for prototrophy in an auxotrophic yeast mutant, for example the URA3, LEU2, LYS2, TRP1, or HIS3 gene.

Since the replication of vectors is conveniently done in *E. coli*, an *E. coli* genetic marker and an *E. coli* origin of replication are advantageously included. These can be obtained from *E. coli* plasmids, such as pBR322, Bluescript® vector or a pUC plasmid, c.g. pUC18 or pUC19, which contain both *E. coli* replication origin and *E. coli* genetic marker conferring resistance to antibiotics, such as ampicillin.

Suitable selectable markers for mammalian cells are those that enable the identification of cells competent to take up nucleic acid binding protein nucleic acid, such as dihydrofolate reductase (DHFR, methotrexate resistance), thymidine kinase, or genes conferring resistance to G418 or hygromycin. The mammalian cell transformants are placed under selection pressure which only those transformants which have taken up and are expressing the marker are uniquely adapted to survive. In the case of a DHFR or glutamine synthase (GS) marker, selection pressure can be imposed by culturing the transformants under conditions in which the pressure is progressively increased, thereby leading to amplification (at its chromosomal integration site) of both the selection gene and the linked DNA that encodes the nucleic acid binding protein. Amplification is the process by which genes in greater demand for the production of a protein critical for growth, together with closely associated genes which may encode a desired protein, are reiterated in tandem within the chromosomes of recombinant cells. Increased quantities of desired protein are usually synthesised from thus amplified DNA.

Expression and cloning vectors usually contain a promoter that is recognised by the host organism and is operably linked to nucleic acid binding protein encoding nucleic acid. Such a promoter may be inducible or constitutive. The promoters are operably

linked to DNA encoding the nucleic acid binding protein by removing the promoter from the source DNA by restriction enzyme digestion and inserting the isolated promoter sequence into the vector. Both the native nucleic acid binding protein promoter sequence and many heterologous promoters may be used to direct  
5 amplification and/or expression of nucleic acid binding protein encoding DNA.

Promoters suitable for use with prokaryotic hosts include, for example, the  $\beta$ -lactamase and lactose promoter systems, alkaline phosphatase, the tryptophan (Trp) promoter system and hybrid promoters such as the tac promoter. Their nucleotide sequences have  
10 been published, thereby enabling the skilled worker operably to ligate them to DNA encoding nucleic acid binding protein, using linkers or adapters to supply any required restriction sites. Promoters for use in bacterial systems will also generally contain a Shine-Delgarno sequence operably linked to the DNA encoding the nucleic acid binding protein.

15

Preferred expression vectors are bacterial expression vectors which comprise a promoter of a bacteriophage such as phage  $\lambda$  or T7 which is capable of functioning in the bacteria. In one of the most widely used expression systems, the nucleic acid encoding the fusion protein may be transcribed from the vector by T7 RNA polymerase  
20 (Studier et al, Methods in Enzymol. 185; 60-89, 1990). In the *E. coli* BL21(DE3) host strain, used in conjunction with pET vectors, the T7 RNA polymerase is produced from the  $\lambda$ -lysogen DE3 in the host bacterium, and its expression is under the control of the IPTG inducible lac UV5 promoter. This system has been employed successfully for over-production of many proteins. Alternatively the polymerase gene may be  
25 introduced on a lambda phage by infection with an int- phage such as the CE6 phage which is commercially available (Novagen, Madison, USA). other vectors include vectors containing the lambda PL promoter such as PLEX (Invitrogen, NL) , vectors containing the trc promoters such as pTrcHisXpressTm (Invitrogen) or pTrc99 (Pharmacia Biotech, SE) or vectors containing the tac promoter such as pKK223-3  
30 (Pharmacia Biotech) or PMAL (New England Biolabs, MA, USA).



Moreover, the nucleic acid binding protein gene according to the invention preferably includes a secretion sequence in order to facilitate secretion of the polypeptide from bacterial hosts, such that it will be produced as a soluble native peptide rather than in an inclusion body. The peptide may be recovered from the bacterial periplasmic space, or the culture medium, as appropriate.

Suitable promoting sequences for use with yeast hosts may be regulated or constitutive and are preferably derived from a highly expressed yeast gene, especially a *Saccharomyces cerevisiae* gene. Thus, the promoter of the TRP1 gene, the ADHI or ADHII gene, the acid phosphatase (PH05) gene, a promoter of the yeast mating pheromone genes coding for the  $\alpha$ - or  $\alpha$ -factor or a promoter derived from a gene encoding a glycolytic enzyme such as the promoter of the enolase, glyceraldehyde-3-phosphate dehydrogenase (GAP), 3-phospho glycerate kinase (PGK), hexokinase, pyruvate decarboxylase, phosphofructokinase, glucose-6-phosphate isomerase, 3-phosphoglycerate mutase, pyruvate kinase, triose phosphate isomerase, phosphoglucose isomerase or glucokinase genes, or a promoter from the TATA binding protein (TBP) gene can be used. Furthermore, it is possible to use hybrid promoters comprising upstream activation sequences (UAS) of one yeast gene and downstream promoter elements including a functional TATA box of another yeast gene, for example a hybrid promoter including the UAS(s) of the yeast PH05 gene and downstream promoter elements including a functional TATA box of the yeast GAP gene (PH05-GAP hybrid promoter). A suitable constitutive PH05 promoter is e.g. a shortened acid phosphatase PH05 promoter devoid of the upstream regulatory elements (UAS) such as the PH05 (-173) promoter element starting at nucleotide -173 and ending at nucleotide -9 of the PH05 gene.

Nucleic acid binding protein gene transcription from vectors in mammalian hosts may be controlled by promoters derived from the genomes of viruses such as polyoma virus, adenovirus, fowlpox virus, bovine papilloma virus, avian sarcoma virus, cytomegalovirus (CMV), a retrovirus and Simian Virus 40 (SV40), from heterologous mammalian promoters such as the actin promoter or a very strong promoter, e.g. a

ribosomal protein promoter, and from the promoter normally associated with nucleic acid binding protein sequence, provided such promoters are compatible with the host cell systems.

- 5 Transcription of a DNA encoding nucleic acid binding protein by higher eukaryotes may be increased by inserting an enhancer sequence into the vector. Enhancers are relatively orientation and position independent. Many enhancer sequences are known from mammalian genes (e.g. elastase and globin). However, typically one will employ an enhancer from a eukaryotic cell virus. Examples include the SV40 enhancer on the  
10 late side of the replication origin (bp 100-270) and the CMV early promoter enhancer. The enhancer may be spliced into the vector at a position 5' or 3' to nucleic acid binding protein DNA, but is preferably located at a site 5' from the promoter.

- Advantageously, a eukaryotic expression vector encoding a nucleic acid binding protein  
15 according to the invention may comprise a locus control region (LCR). LCRs are capable of directing high-level integration site independent expression of transgenes integrated into host cell chromatin, which is of importance especially where the nucleic acid binding protein gene is to be expressed in the context of a permanently-transfected eukaryotic cell line in which chromosomal integration of the vector has occurred, or in  
20 transgenic animals.

- Eukaryotic vectors may also contain sequences necessary for the termination of transcription and for stabilising the mRNA. Such sequences are commonly available from the 5' and 3' untranslated regions of eukaryotic or viral DNAs or cDNAs. These  
25 regions contain nucleotide segments transcribed as polyadenylated fragments in the untranslated portion of the mRNA encoding nucleic acid binding protein.

- An expression vector includes any vector capable of expressing nucleic acid binding protein nucleic acids that are operatively linked with regulatory sequences, such as  
30 promoter regions, that are capable of expression of such DNAs. Thus, an expression vector refers to a recombinant DNA or RNA construct, such as a plasmid, a phage,

recombinant virus or other vector, that upon introduction into an appropriate host cell, results in expression of the cloned DNA. Appropriate expression vectors are well known to those with ordinary skill in the art and include those that are replicable in eukaryotic and/or prokaryotic cells and those that remain episomal or those which  
5 integrate into the host cell genome. For example, DNAs encoding nucleic acid binding protein may be inserted into a vector suitable for expression of cDNAs in mammalian cells, e.g. a CMV enhancer-based vector such as pEVRF (Matthias, et al., (1989) NAR 17, 6418).

10 Particularly useful for practising the present invention are expression vectors that provide for the transient expression of DNA encoding nucleic acid binding protein in mammalian cells. Transient expression usually involves the use of an expression vector that is able to replicate efficiently in a host cell, such that the host cell accumulates many copies of the expression vector, and, in turn, synthesises high levels of nucleic  
15 acid binding protein. For the purposes of the present invention, transient expression systems are useful e.g. for identifying nucleic acid binding protein mutants, to identify potential phosphorylation sites, or to characterise functional domains of the protein.

Construction of vectors according to the invention employs conventional ligation  
20 techniques. Isolated plasmids or DNA fragments are cleaved, tailored, and religated in the form desired to generate the plasmids required. If desired, analysis to confirm correct sequences in the constructed plasmids is performed in a known fashion. Suitable methods for constructing expression vectors, preparing in vitro transcripts, introducing DNA into host cells, and performing analyses for assessing nucleic acid binding protein  
25 expression and function are known to those skilled in the art. Gene presence, amplification and/or expression may be measured in a sample directly, for example, by conventional Southern blotting, Northern blotting to quantitate the transcription of mRNA, dot blotting (DNA or RNA analysis), or in situ hybridisation, using an appropriately labelled probe which may be based on a sequence provided herein. Those  
30 skilled in the art will readily envisage how these methods may be modified, if desired.

In accordance with another embodiment of the present invention, there are provided cells containing the above-described nucleic acids. Such host cells such as prokaryote, yeast and higher eukaryote cells may be used for replicating DNA and producing the nucleic acid binding protein. Suitable prokaryotes include eubacteria, such as Gram-negative or Gram-positive organisms, such as *E. coli*, e.g. *E. coli* K-12 strains, DH5a and HB101, or Bacilli. Further hosts suitable for the nucleic acid binding protein encoding vectors include eukaryotic microbes such as filamentous fungi or yeast, e.g. *Saccharomyces cerevisiae*. Higher eukaryotic cells include insect and vertebrate cells, particularly mammalian cells including human cells or nucleated cells from other multicellular organisms. In recent years propagation of vertebrate cells in culture (tissue culture) has become a routine procedure. Examples of useful mammalian host cell lines are epithelial or fibroblastic cell lines such as Chinese hamster ovary (CHO) cells, NIH 3T3 cells, HeLa cells or 293T cells. The host cells referred to in this disclosure comprise cells in *in vitro* culture as well as cells that are within a host animal.

DNA may be stably incorporated into cells or may be transiently expressed using methods known in the art. Stably transfected mammalian cells may be prepared by transfecting cells with an expression vector having a selectable marker gene, and growing the transfected cells under conditions selective for cells expressing the marker gene. To prepare transient transfectants, mammalian cells are transfected with a reporter gene to monitor transfection efficiency.

To produce such stably or transiently transfected cells, the cells should be transfected with a sufficient amount of the nucleic acid binding protein-encoding nucleic acid to form the nucleic acid binding protein. The precise amounts of DNA encoding the nucleic acid binding protein may be empirically determined and optimised for a particular cell and assay.

Host cells are transfected or, preferably, transformed with the above-captioned expression or cloning vectors of this invention and cultured in conventional nutrient

media modified as appropriate for inducing promoters, selecting transformants, or amplifying the genes encoding the desired sequences. Heterologous DNA may be introduced into host cells by any method known in the art, such as transfection with a vector encoding a heterologous DNA by the calcium phosphate coprecipitation technique or by electroporation. Numerous methods of transfection are known to the skilled worker in the field. Successful transfection is generally recognised when any indication of the operation of this vector occurs in the host cell. Transformation is achieved using standard techniques appropriate to the particular host cells used.

- 10 Incorporation of cloned DNA into a suitable expression vector, transfection of eukaryotic cells with a plasmid vector or a combination of plasmid vectors, each encoding one or more distinct genes or with linear DNA, and selection of transfected cells are well known in the art (see, e.g. Sambrook et al. (1989) *Molecular Cloning: A Laboratory Manual*, Second Edition, Cold Spring Harbor Laboratory Press).

15

Transfected or transformed cells are cultured using media and culturing methods known in the art, preferably under conditions, whereby the nucleic acid binding protein encoded by the DNA is expressed. The composition of suitable media is known to those in the art, so that they can be readily prepared. Suitable culturing media are also

- 20 commercially available.

Nucleic acid binding proteins according to the invention may be employed in a wide variety of applications, including diagnostics and as research tools. Advantageously, they may be employed as diagnostic tools for identifying the presence of nucleic acid molecules in a complex mixture. nucleic acid binding molecules according to the invention can differentiate single base pair changes in target nucleic acid molecules.

25

Accordingly, the invention provides a method for determining the presence of a target nucleic acid molecule, comprising the steps of:

30

- a) preparing a nucleic acid binding protein by the method set forth above which is specific for the target nucleic acid molecule;
- b) exposing a test system comprising the target nucleic acid molecule to the nucleic acid binding protein under conditions which promote binding, and removing any nucleic acid binding protein which remains unbound;
- c) detecting the presence of the nucleic acid binding protein in the test system.

In a preferred embodiment, the nucleic acid binding molecules of the invention can be incorporated into an ELISA assay. For example, phage displaying the molecules of the invention can be used to detect the presence of the target nucleic acid, and visualised using enzyme-linked anti-phage antibodies.

Further improvements to the use of zinc finger phage for diagnosis can be made, for example, by co-expressing a marker protein fused to the minor coat protein (gVIII) of bacteriophage. Since detection with an anti-phage antibody would then be obsolete, the time and cost of each diagnosis would be further reduced. Depending on the requirements, suitable markers for display might include the fluorescent proteins (A. B. Cubitt, *et al.*, (1995) *Trends Biochem Sci.* **20**, 448-455; T. T. Yang, *et al.*, (1996) *Gene* **173**, 19-23), or an enzyme such as alkaline phosphatase which has been previously displayed on gIII (J. McCafferty, R. H. Jackson, D. J. Chiswell, (1991) *Protein Engineering* **4**, 955-961). Labelling different types of diagnostic phage with distinct markers would allow multiplex screening of a single nucleic acid sample. Nevertheless, even in the absence of such refinements, the basic ELISA technique is reliable, fast, simple and particularly inexpensive. Moreover it requires no specialised apparatus, nor does it employ hazardous reagents such as radioactive isotopes, making it amenable to routine use in the clinic. The major advantage of the protocol is that it obviates the requirement for gel electrophoresis, and so opens the way to automated nucleic acid diagnosis.

The invention provides nucleic acid binding proteins which can be engineered with exquisite specificity. The invention lends itself, therefore, to the design of any

molecule of which specific nucleic acid binding is required. For example, the proteins according to the invention may be employed in the manufacture of chimeric restriction enzymes, in which a nucleic acid cleaving domain is fused to a nucleic acid binding domain comprising a zinc finger as described herein.

5

Moreover, the invention provides therapeutic agents and methods of therapy involving use of nucleic acid binding proteins as described herein. In particular, the invention provides the use of polypeptide fusions comprising an integrase, such as a viral integrase, and a nucleic acid binding protein according to the invention to target nucleic acid sequences *in vivo* (Bushman, (1994) PNAS (USA) 91:9233-9237). In gene therapy applications, the method may be applied to the delivery of functional genes into defective genes, or the delivery of nonsense nucleic acid in order to disrupt undesired nucleic acid. Alternatively, genes may be delivered to known, repetitive stretches of nucleic acid, such as centromeres, together with an activating sequence such as an LCR. This would represent a route to the safe and predictable incorporation of nucleic acid into the genome.

In conventional therapeutic applications, nucleic acid binding proteins according to the invention may be used to specifically knock out cell having mutant vital proteins. For example, if cells with mutant ras are targeted, they will be destroyed because ras is essential to cellular survival. Alternatively, the action of transcription factors may be modulated, preferably reduced, by administering to the cell agents which bind to the binding site specific for the transcription factor. For example, the activity of HIV tat may be reduced by binding proteins specific for HIV TAR.

25

Moreover, binding proteins according to the invention may be coupled to toxic molecules, such as nucleases, which are capable of causing irreversible nucleic acid damage and cell death. Such agents are capable of selectively destroying cells which comprise a mutation in their endogenous nucleic acid.

30

Nucleic acid binding proteins and derivatives thereof as set forth above may also be applied to the treatment of infections and the like in the form of organism-specific antibiotic or antiviral drugs. In such applications, the binding proteins may be coupled to a nuclease or other nuclear toxin and targeted specifically to the nucleic acids of  
5 microorganisms.

The invention likewise relates to pharmaceutical preparations which contain the compounds according to the invention or pharmaceutically acceptable salts thereof as active ingredients, and to processes for their preparation.

10

The pharmaceutical preparations according to the invention which contain the compound according to the invention or pharmaceutically acceptable salts thereof are those for enteral, such as oral, furthermore rectal, and parenteral administration to (a) warm-blooded animal(s), the pharmacological active ingredient being present on its own  
15 or together with a pharmaceutically acceptable carrier. The daily dose of the active ingredient depends on the age and the individual condition and also on the manner of administration.

The novel pharmaceutical preparations contain, for example, from about 10 % to about  
20 80%, preferably from about 20 % to about 60 %, of the active ingredient. Pharmaceutical preparations according to the invention for enteral or parenteral administration are, for example, those in unit dose forms, such as sugar-coated tablets, tablets, capsules or suppositories, and furthermore ampoules. These are prepared in a manner known per se, for example by means of conventional mixing, granulating,  
25 sugar-coating, dissolving or lyophilising processes. Thus, pharmaceutical preparations for oral use can be obtained by combining the active ingredient with solid carriers, if desired granulating a mixture obtained, and processing the mixture or granules, if desired or necessary, after addition of suitable excipients to give tablets or sugar-coated tablet cores.

30



- Suitable carriers are, in particular, fillers, such as sugars, for example lactose, sucrose, mannitol or sorbitol, cellulose preparations and/or calcium phosphates, for example tricalcium phosphate or calcium hydrogen phosphate, furthermore binders, such as starch paste, using, for example, corn, wheat, rice or potato starch, gelatin, tragacanth, methylcellulose and/or polyvinylpyrrolidone, if desired, disintegrants, such as the abovementioned starches, furthermore carboxymethyl starch, crosslinked polyvinylpyrrolidone, agar, alginic acid or a salt thereof, such as sodium alginate; auxiliaries are primarily glidants, flow-regulators and lubricants, for example silicic acid, talc, stearic acid or salts thereof, such as magnesium or calcium stearate, and/or polyethylene glycol. Sugar-coated tablet cores are provided with suitable coatings which, if desired, are resistant to gastric juice, using, inter alia, concentrated sugar solutions which, if desired, contain gum arabic, talc, polyvinylpyrrolidone, polyethylene glycol and/or titanium dioxide, coating solutions in suitable organic solvents or solvent mixtures or, for the preparation of gastric juice-resistant coatings, solutions of suitable cellulose preparations, such as acetylcellulose phthalate or hydroxypropylmethylcellulose phthalate. Colorants or pigments, for example to identify or to indicate different doses of active ingredient, may be added to the tablets or sugar-coated tablet coatings.
- Other orally utilisable pharmaceutical preparations are hard gelatin capsules, and also soft closed capsules made of gelatin and a plasticiser, such as glycerol or sorbitol. The hard gelatin capsules may contain the active ingredient in the form of granules, for example in a mixture with fillers, such as lactose, binders, such as starches, and/or lubricants, such as talc or magnesium stearate, and, if desired, stabilisers. In soft capsules, the active ingredient is preferably dissolved or suspended in suitable liquids, such as fatty oils, paraffin oil or liquid polyethylene glycols, it also being possible to add stabilisers.

- Suitable rectally utilisable pharmaceutical preparations are, for example, suppositories, which consist of a combination of the active ingredient with a suppository base. Suitable suppository bases are, for example, natural or synthetic triglycerides, paraffin

hydrocarbons, polyethylene glycols or higher alkanols. Furthermore, gelatin rectal capsules which contain a combination of the active ingredient with a base substance may also be used. Suitable base substances are, for example, liquid triglycerides, polyethylene glycols or paraffin hydrocarbons.

5

Suitable preparations for parenteral administration are primarily aqueous solutions of an active ingredient in water-soluble form, for example a water-soluble salt, and furthermore suspensions of the active ingredient, such as appropriate oily injection suspensions, using suitable lipophilic solvents or vehicles, such as fatty oils, for example sesame oil, or synthetic fatty acid esters, for example ethyl oleate or triglycerides, or aqueous injection suspensions which contain viscosity-increasing substances, for example sodium carboxymethylcellulose, sorbitol and/or dextran, and, if necessary, also stabilisers.

15 The dose of the active ingredient depends on the warm-blooded animal species, the age and the individual condition and on the manner of administration. In the normal case, an approximate daily dose of about 10 mg to about 250 mg is to be estimated in the case of oral administration for a patient weighing approximately 75 kg .

20 The invention is described below, for the purpose of illustration only, in the following examples.

### **Example 1**

#### **Determination of binding site preferences in zinc fingers**

25

#### *Design Of Zinc Finger Phage Display Libraries*

Zinc finger-DNA recognition at the interface between adjacent DNA subsites is studied using a zinc finger phage display library. This library is based on the three-finger DNA-binding domain of Zif268, but contains randomisations of amino acids from finger 2 (F2) and finger 3 (F3), at residue positions which could form a network of contacts across the interface of their DNA subsites. The detailed design of the library

30

is shown in Figure 1c, together with the generic DNA binding site used in selections. Briefly, the library contains randomisations at F2 residue position 6 (hereafter denoted F2[+6]) and F3 residue positions -1, +1, +2 and +3 (hereafter denoted F3[-1], F3[+2], etc.).

5

Library selections are carried out using DNA binding sites that resembled the Zif268 operator, but which contained systematic combinations of bases in the DNA doublet which forms the base-step between the DNA subsites of F2 and F3. DNA binding sites are of the generic form 5'-GNX-XCG-GCG-3', where X-X denotes a given  
10 combination of the bases at the interface between the DNA subsites, and N denotes that the four bases are equally represented at DNA position 3. Thus the interaction between F3[+3] and nucleotide position 3N is allowed complete freedom in this experiment. This feature of the library allows selection of a large family (or database) of related zinc fingers that bind a given combination of bases at nucleotide positions 4X and 5X,  
15 but which are non-identical owing to different interactions with the middle base in the nominal triplet subsite of F3.

The first library to be constructed, LIB-A, contains randomisations at F2 residue position 6 and F3 residue positions -1, 1, 2 and 3 (see Figure 2), and is sorted using the  
20 DNA sequence 5'-GNX-XCG-GCG-3', where X-X denotes a known combination of the two bases at DNA positions 4X and 5X, and N denotes an equal probability of any of the four bases at DNA position 3. The second library, LIB-B, contains randomisations at F2 residue position 6 and F3 residue positions -1 and 2, and is sorted using the DNA sequence 5'-GCX-XCG-GCG-3', where X-X denotes a known combination of the two  
25 bases at DNA positions 4X and 5X.

The genes for the two different zinc finger phage display libraries are assembled from four synthetic DNA oligonucleotides by directional end-to-end ligation using three short complementary DNA linkers. The oligonucleotides contain selectively randomised  
30 codons (of sequence NNS; N = A/C/G/T, S = G/C) in the appropriate amino acid positions of fingers 2 and 3. The constructs are amplified by PCR using primers

containing *Not I* and *Sfi I* restriction sites, digested with the above endonucleases to produce cloning overhangs, and ligated into phage vector Fd-Tet-SN. Electrocompetent *E. coli* TG 1 cells are transformed with the recombinant vector and plated onto TYE medium (1.5% agar, 1% Bacto tryptone, 0.5% Bacto yeast extract, 5 0.8% NaCl) containing 15 µg/ml tetracycline.

Allowing this freedom to some protein-DNA interactions that are not being studied is a useful strategy towards increasing the diversity of clones which can be obtained from any one selection experiment. However, at the same time, it is important to limit the 10 number of contacts that are allowed contextual freedom at any one time, otherwise there is a danger that a subset of particularly strong intermolecular interactions will dominate the selections. Anticipating this eventuality, a smaller sublibrary is also created that contains randomised residues only in positions F2[+6] and F3[-1 and +2], and therefore does not allow for contextual freedom in selections. Clones selected from 15 this library are marked with an asterisk when they are discussed herein.

#### *Experimental Strategy*

Phage selections from the two zinc finger libraries are performed separately in order to determine the diversity of DNA sequences which can be bound specifically by members 20 of each library. Sixteen selections are performed on each library, using the different DNA binding sites that correspond to all 16 possible combinations of bases at nucleotide positions 4X and 5X. The DNA binding site used to select specifically binding phage is immobilised on a solid surface, while a 10-fold excess of each of the other 15 DNA sites is present in solution as a specific competitor.

25

#### *Phage Selections*

Tetracycline resistant colonies are transferred from plates into 2xTY medium (16g/litre Bacto tryptone, 10g/litre Bacto yeast extract, 5g/litre NaCl) containing 50µM ZnCl<sub>2</sub> and 15 µg/ml tetracycline, and cultured overnight at 30°C in a shaking incubator. 30 Cleared culture supernatant containing phage particles is obtained by centrifuging at 300g for 5 minutes.

Biotinylated DNA target sites (1pmol) are bound to streptavidin-coated tubes (Boehringer Mannheim). Phage supernatant solutions are diluted 1:10 in PBS selection buffer (PBS containing 50µM ZnCl<sub>2</sub>, 2% Marvel, 1% Tween, 20µg/ml sonicated salmon sperm DNA, 10 pmol/ml of each of the 15 other possible unbiotinylated DNA sites), and 1 ml is applied to each tube for 1 hour at 20°C. After this time, the tubes are emptied and washed 20 times with PBS containing 50µM ZnCl<sub>2</sub>, 2% Marvel and 1% Tween. Retained phage are eluted in 0.1ml 0.1M triethylamine and neutralised with an equal volume of 1M Tris (pH 7.4). Logarithmic-phase *E. coli* TG 1 (0.5ml) are infected with eluted phage (50µl), and used to prepare phage supernatants for subsequent rounds of selection. After 3 rounds of selection, *E. coli* infected with selected phage are plated, individual colonies are picked and used to grow phage for binding site signature assays and DNA sequencing.

After three rounds of phage selection against a particular DNA binding site, individual zinc finger clones are recovered, and the DNA binding specificity of each clone is determined by the binding site signature method. This involves screening each zinc finger phage for binding to eight different libraries of the DNA binding site, designed such that each library contains one fixed base and one randomised base at either of positions 4X and 5X (i.e. libraries GN, AN, TN, CN, and NG, NA, NT, NC). Thus each of the 16 DNA binding sites used in selection experiments is specified by a unique combination of two libraries - for example, the DNA binding site containing 4G5G is present in only two of the eight libraries in which the relevant doublet had one nucleotide randomised and the other nucleotide fixed as guanine, i.e. libraries 4G5N and 4N5G. The eight DNA libraries used in binding site signatures are arrayed across a microtitre plate and zinc finger phage binding is detected by phage ELISA. The pattern of binding to the eight DNA libraries reveals the DNA sequence specificity (or preference) of each phage clone, and only those clones found to be relatively specific are subsequently sequenced to reveal the identity of the amino acids present in the randomised zinc finger residue positions.

Procedures are as described previously (Choo, Y. & Klug, A. (1994) Proc. Natl. Acad. Sci. USA 91, 11163-11167; Choo, Y. & Klug, A. (1994) Proc. Natl. Acad. Sci. USA 91, 11168-11172). Briefly, 5'-biotinylated positionally randomised oligonucleotide libraries, containing Zif268 operator variants, are synthesised by primer extension as described. DNA libraries (0.4 pmol/well for LIB-A and 1.2 pmol/well for LIB-B) are added to streptavidin-coated ELISA wells (Boehringer-Mannheim) in PBS containing 50µM ZnCl<sub>2</sub> (PBS/Zn). Phage solution (overnight bacterial culture supernatant diluted 1:10 in PBS/Zn containing 2% Marvel, 1% Tween and 20µg/ml sonicated salmon sperm DNA) are applied to each well (50µl/well). Binding is allowed to proceed for one hour at 20°C. Unbound phage are removed by washing 6 times with PBS/Zn containing 1% Tween, then 3 times with PBS/Zn. Bound phage are detected by ELISA using horseradish peroxidase-conjugated anti-M13 IgG (Pharmacia Biotech) and the colourimetric signal quantitated using SOFFMAX 2.32 (Molecular Devices).

The coding sequence of individual zinc finger clones is amplified by PCR using external primers complementary to phage sequence. These PCR products are sequenced manually using Thermo Sequenase cycle sequencing (Amersham Life Science).

#### *Analysis Of Phage-Selected Zinc Fingers*

Figure 3 shows the binding site signatures of relatively sequence-specific zinc finger phages selected from both libraries, using the 16 different DNA doublets which form the base-step between the DNA subsites of fingers 2 and 3. The results show that zinc finger clones are selected which bind specifically to almost all subsites, including those triplets in which the 5' position (nucleotide 5X in the model system) is fixed as a base other than guanine. Overall, the selections show that any of the four bases can be bound specifically in both the 5' and 3' positions of a nominal triplet subsite. The results are summarised in Figure 4.

Selections from the smaller sub-library yield fingers that can bind specifically to only 8 of the 16 doublets, whereas members of the larger library yield fingers that recognise

15 out of the 16 doublets. It is not known whether this difference in efficacy originates from the inclusion of more randomised positions in the larger library, or the conformational flexibility afforded by the contextual freedom designed into the larger library, or both. The only base-step that does not yield specific zinc fingers is 4G5A.

5 This dinucleotide may induce an unfavourable DNA deformation in the context of the DNA binding sites used for selection.

### **Example 2**

#### **Determination of +2 specificity for position 1**

10

The amino acid present in  $\alpha$ -helical position 2 of a zinc finger can help determine the specificity for the base-pair at the interface of two overlapping DNA quadruplet subsites (see Figure 1B; position 5/5', corresponding to position 1 or 4 of the quadruplet as discussed above). An Asp residue present in F3[+2] of wild-type Zif268

15 has been shown to play a role in DNA recognition, and further examples are generated by the current phage display experiments (See Example 1 for details, and Figure 5A).

The experimental protocol followed is that of Example 1. Figure 5A shows an example of related zinc finger clones showing the effect of  $\alpha$ -helical position 2 on DNA-binding

20 specificity. In this case, position 6 of finger 2 is invariant (Asn) and the change in case specificity in the zinc finger in order to select for contact to this base is dictated by position +2 in finger 3.

This family of zinc fingers is derived from selections using DNA binding sites

25 containing 4T5A or 4T5C subsite interfaces. The base preference for the 5X- 5'X base-pair is determined by the amino acid present at F3[+2], probably by the formation of cross-strand contacts.

Figure 5B shows examples of correlations between certain amino acids selected at

30 F3[+2] and the identity of the base present at position 5'X. Selections reveal the possibility of DNA contacts from five amino acids (Asn, Gln, Arg, Lys and His) which

are all capable of donating a H-bond to the exocyclic oxygen atom of either guanine (O<sub>6</sub>) or thymine (O<sub>4</sub>) in nucleotide position 5'X. The clones isolated with these amino acids at F3[+2] are listed in this diagram together with the binding site signature showing the base-preference at position 5'X. Overall, Ser dominated the selections with an occurrence of 38%, in accord with its presence in position 2 in over half of all known zinc fingers. Threonine, Ala and Gly occurred frequently in the selections (15%, 15% and 9% respectively) but did not show any discernible patterns of discrimination. Certain amino acids (Cys, Asp, Phe, Ile, Leu, Met, Pro, Val and Trp) are never selected in position 2. Their ability to bind in certain situations is however not to be excluded.

A small subset of amino acids selected in F3[+2] show significant correlations to the identity of the base-pair in position 5'X (Figure 5B), suggesting that cross-strand interactions between these may be a general mechanism of DNA-recognition. Most of these correlations can be rationalised as pairings between hydrogen bond donors in F3[+2] and guanine or thymine in DNA position 5'X, in accordance with the framework of the Zif268 model. In contrast to amino acids that are never selected in position 2, or amino acids that are selected but which show no significant correlations, the amino acids which consistently appear to play a role in DNA recognition from this position have side chains with multiple hydrogen bonding groups. It is possible that these residues can play a role in base recognition because they achieve greater specificity by participating in buttressing networks.

### **Example 3**

#### **Construction of a General Purpose Library**

The binary library system constructed in this example comprises libraries LIB1/2 and LIB2/3 that each encode the three fingers of Zif268 but with some amino acid positions selectively randomised. Instead of adhering to the model of modular zinc fingers, the new libraries contain concerted variations in certain amino acid positions in adjacent zinc fingers. Thus LIB1/2 contains simultaneous variations in F1 positions -1, 2, 3, 5



and 6 and F2 positions -1, 1, 2 and 3. LIB2/3 contains simultaneous variations in F2 positions 3 and 6 and F3 positions -1, 1, 2, 3 and 5, 6. The remaining amino acids in each library are as the WT Zif268 sequence. The two libraries are cloned in Fd phage as GIII fusions according to standard protocols.

5

The amino acids that are allowed at each varied position are as follows:

#### LIB1/2

- F1 pos. -1 = R, Q, H, N, D, A, T;  
 pos. 2 = D, A, R, Q, H, K, S, N;  
 10 pos. 3 = H, N, S, T, V, A, D;  
 pos. 5 = I, T;  
 pos. 6 = R, Q, V, A, E, K, N, T.  
 F2 pos. -1 = R, Q, H, N, D, A, T;  
 pos. 1 = S, R;  
 15 pos. 2 = D, A, R, Q, H, K, S, N;  
 pos. 3 = H, N, S, T, V, A, D;

#### LIB2/3

- F2 pos. 3 = H, N, S, T, V, A, D;  
 20 pos. 6 = R, Q, V, A, E, K, N, T.  
 F3 pos. -1 = R, Q, H, N, D, A, T;  
 pos. 1 = R, K, S, N;  
 pos. 2 = D, A, R, Q, H, K, S, N;  
 pos. 3 = H, N, S, T, V, A, D;  
 25 pos. 5 = K, I, T;  
 pos. 6 = R, Q, V, A, E, K, N, T.

#### *Selections And Recombinations*

- Selections are performed using the DNA sequence GCG-GMN-OPQ for LIB 1/2 and  
 30 the DNA sequence IJK-LMG-GCG for LIB2/3, where the underlined bases are bound  
 by the WT Zif268 residues and each of the other letters stands for any given nucleotide.  
 The conserved nucleotides of the Zif268 binding site serve to fix the register of the

interaction by binding to the conserved portion of the Zif268 DNA-binding domain. The binary phage display libraries can be mixed so that selections using these two generic sites are performed in a single tube, or the selections can be performed separately. After a number of rounds of selection the two libraries are recombined to  
5 produce a chimaeric DNA-binding domain that recognises the sequence IJK-LMN-OPQ.

The recombination reactions are performed by amplifying the selected three-finger domains by PCR and cutting the PCR products using restriction enzyme Ddel. This  
10 cuts the genes of both zinc finger libraries at the DNA sequence coding for F2  $\alpha$ -helical positions 4 and 5. The digested products are randomly religated to produce recombinant genes coding for the chimaeric DNA-binding domains (and other products including reconstituted WT Zif268). The chimaeric DNA-binding domains are selectively amplified from the mixture of products by PCR using selective primers that recognise  
15 the recombinant F1 and F3 genes, rather than WT genes, and cloned in Fd phage (for more selections) or other vectors (e.g. for expression in *E coli*).

The initial selections from the binary libraries can be pushed to completion, thus allowing the assembly of a single clone by recombination. Alternatively, if the initial  
20 selections are less stringent, many candidates will be available for the assembly of various chimaeric domains after recombination. In the latter case, the best recombinant protein can be selected by further rounds of selection on phage.

**Claims**

1. A zinc finger polypeptide library in which each polypeptide comprises more than one zinc finger which has been at least partially randomised.
- 5 2. A library according to claim 1 wherein two zinc fingers are at least partially randomised in each polypeptide.
3. A library according to claim 1 or claim 2, wherein the randomised zinc fingers  
10 are adjacent.
4. A set of zinc finger polypeptide libraries which encode overlapping zinc finger polypeptides, each polypeptide comprising more than one zinc finger which has been at least partially randomised, and which polypeptides may be assembled after selection to  
15 form a multifinger zinc finger polypeptide.
5. A set according to claim 4, comprising a pair of libraries encoding three-zinc finger polypeptides.
- 20 6. A library or set of libraries according to any preceding claim, wherein the randomised positions are selected from positions -1, 1, 2, 3, 5 and 6.
7. A library according to any preceding claim, wherein the randomisation of amino acid residues is restricted such that the following amino acids may appear at the given  
25 positions:

Position	Possible Amino Acids
-1	R, Q, H, N, D, A, T
1	S, R, K, N
2	D, A, R, Q, H, K, S, N
3	H, N, S, T, V, A, D
5	I, T, K

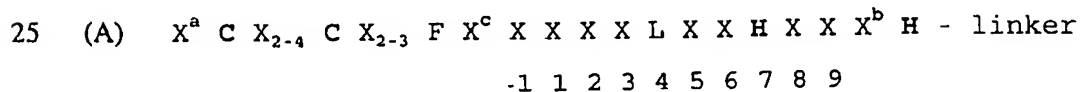
8. A set of two libraries according to claim 7 for selecting a three-finger zinc finger protein, wherein the following amino acids may appear at the given positions:

Library 1		Library 2	
F1:	amino acid	F1:	amino acid
-1	R, Q, H, N, D, A		
2	D, A, R, Q, H, K, S, N		
3	H, N, S, T, V, A, D		
5	I, T		
6	R, Q, V, A, E, K, N, T		
F2			
-1	R, Q, H, N, D, A, T		
1	S, R		
2	D, A, R, Q, H, K, S, N		
3	H, N, S, T, V, A, D	3	H, N, S, T, V, A, D
		6	R, Q, V, A, E, K, N, T
F3			
		-1	R, Q, H, N, D, A, T
		1	R, K, S, N
		2	D, A, R, Q, H, K, S, N
		3	H, N, S, T, V, A, D
		5	K, I, T
		6	R, Q, V, A, E, K, N, T

9. A library according to claim 1, wherein the amino acids at positions -1, 2, 3 and 6 are selected as follows:

- a) if base 4 in the quadruplet is G, then position +6 in the  $\alpha$ -helix is Arg or Lys;
- b) if base 4 in the quadruplet is A, then position +6 in the  $\alpha$ -helix is Glu, Asn or Val;
- c) if base 4 in the quadruplet is T, then position +6 in the  $\alpha$ -helix is Ser, Thr, Val or Lys;
- 5 d) if base 4 in the quadruplet is C, then position +6 in the  $\alpha$ -helix is Ser, Thr, Val, Ala, Glu or Asn;
- e) if base 3 in the quadruplet is G, then position +3 in the  $\alpha$ -helix is His;
- f) if base 3 in the quadruplet is A, then position +3 in the  $\alpha$ -helix is Asn;
- g) if base 3 in the quadruplet is T, then position +3 in the  $\alpha$ -helix is Ala, Ser or Val;
- 10 provided that if it is Ala, then one of the residues at -1 or +6 is a small residue;
- h) if base 3 in the quadruplet is C, then position +3 in the  $\alpha$ -helix is Ser, Asp, Glu, Leu, Thr or Val;
- i) if base 2 in the quadruplet is G, then position -1 in the  $\alpha$ -helix is Arg;
- j) if base 2 in the quadruplet is A, then position -1 in the  $\alpha$ -helix is Gln;
- 15 k) if base 2 in the quadruplet is T, then position -1 in the  $\alpha$ -helix is His or Thr;
- l) if base 2 in the quadruplet is C, then position -1 in the  $\alpha$ -helix is Asp or His.
- m) if base 1 in the quadruplet is G, then position +2 is Glu;
- n) if base 1 in the quadruplet is A, then position +2 Arg or Gln;
- o) if base 1 in the quadruplet is C, then position +2 is Asn, Gln, Arg, His or Lys;
- 20 if base 1 in the quadruplet is T, then position +2 is Ser or Thr.

10. A library according to any preceding claim, wherein each zinc finger has the general primary structure



wherein X (including X<sup>a</sup>, X<sup>b</sup> and X<sup>c</sup>) is any amino acid.

- 30 11. A library according to claim 10 wherein  $X^a$  is  $F/Y-X$  or  $P-F/Y-X$ .

12. A library according to claim 10 or claim 11 wherein  $X_{2,4}$  is selected from any one of: S-X, E-X, K-X, T-X, P-X and R-X.
13. A library according to any one of claims 10 to 12 wherein  $X^b$  is T or I.
- 5 14. A library according to any one of claims 10 to 13 wherein  $X_{2,3}$  is G-K-A, G-K-C, G-K-S, G-K-G, M-R-N or M-R.
15. A library according to any one of claims 10 to 14 wherein the linker is T-G-E-K  
10 or T-G-E-K-P.
16. A library according to any one of claims 10 to 15 wherein position +9 is R or K.
- 15 17. A library according to any one of claims 10 to 16 wherein positions +1, +5 and +8 are not occupied by any one of the hydrophobic amino acids, F, W or Y.
18. A library according to claim 17 wherein positions +1, +5 and +8 are occupied by the residues K, T and Q respectively.
- 20 19. A method for preparing a library of nucleic acid binding proteins of the Cys2-His2 zinc finger class capable of binding to a target nucleic acid sequence, comprising the steps of:
- 25 a) selecting a model zinc finger polypeptide from the group consisting of naturally occurring zinc finger polypeptides and consensus zinc finger polypeptides; and
- b) randomising more than one finger therein according to any one of claims 1 to 9.
- 30 20. A method according to claim 19, wherein the model zinc finger is a consensus zinc finger whose structure is selected from the group consisting of the consensus

structure P Y K C P E C G K S F S Q K S D L V K H Q R T H T G, and the consensus structure P Y K C S E C G K A F S Q K S N L T R H Q R I H T G E K P.

21. A method according to claim 19 wherein the model zinc finger is a naturally occurring zinc finger whose structure is selected from one finger of a protein selected from the group consisting of Zif 268 (Elrod-Erickson *et al.*, (1996) Structure 4:1171-1180), GLI (Pavletich and Pabo, (1993) Science 261:1701-1707), Tramtrack (Fairall *et al.*, (1993) Nature 366:483-487) and YY1 (Houbaviy *et al.*, (1996) PNAS (USA) 93:13577-13582).
22. A method according to claim 21 wherein the model zinc finger is finger 2 of Zif 268.
23. A method for determining the presence of a target nucleic acid molecule, comprising the steps of:
- a) preparing a nucleic acid binding protein by the method of any preceding claim which is specific for the target nucleic acid molecule;
  - b) exposing a test system comprising the target nucleic acid molecule to the nucleic acid binding protein under conditions which promote binding, and removing any nucleic acid binding protein which remains unbound;
  - c) detecting the presence of the nucleic acid binding protein in the test system.
24. A method according to claim 23, wherein the presence of the nucleic acid binding protein in the test system is detected by means of an antibody.
25. A method according to claim 23 or claim 24 wherein the nucleic acid binding protein, in use, is displayed on the surface of a filamentous bacteriophage and the presence of the nucleic acid binding protein is detected by detecting the bacteriophage or a component thereof.

1/6

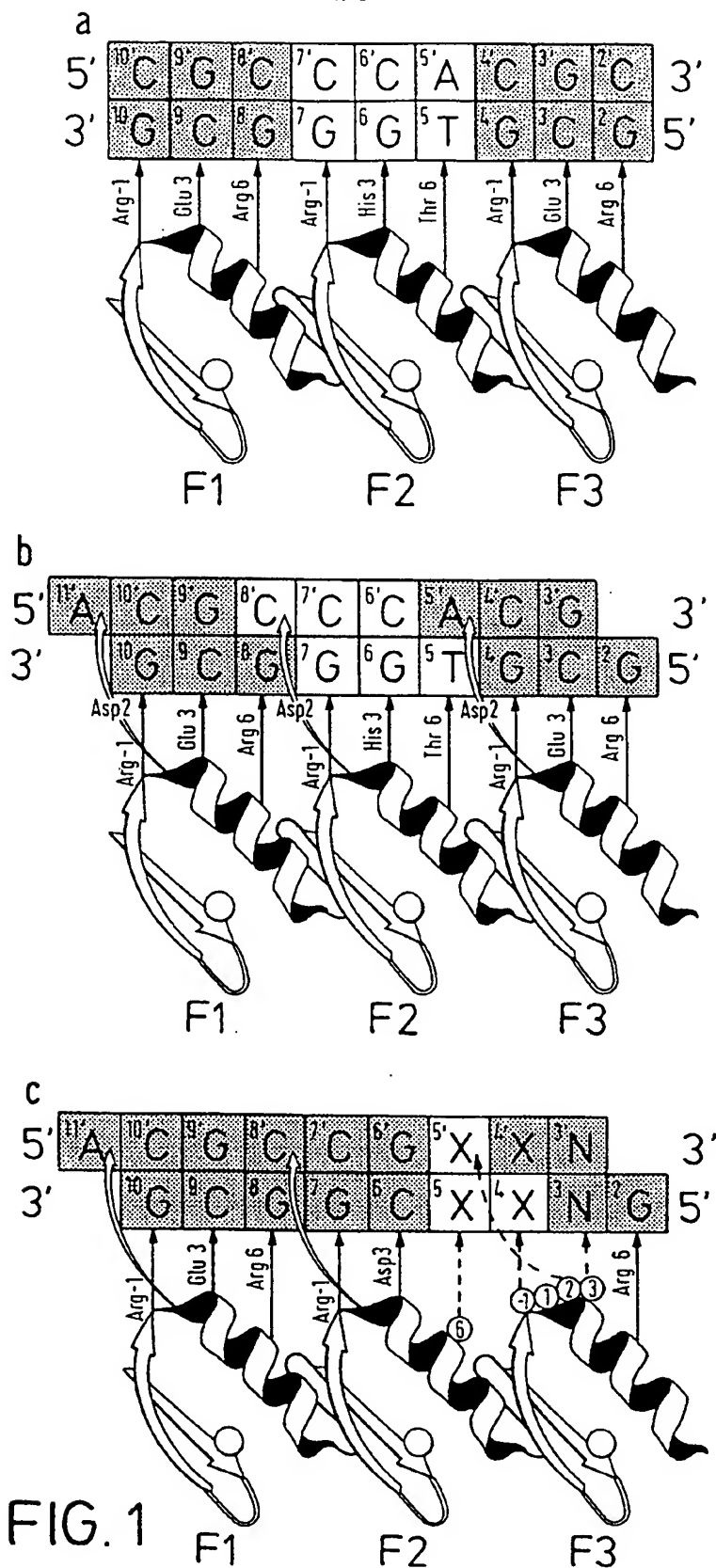


FIG. 1

SUBSTITUTE SHEET (RULE 26)



[illegible]

FIG. 2

3/6

BASE AT POSITIONS 4'X AND 5'X OF DNA SITES USED IN SELECTIONS	AMINO ACIDS SELECTED IN FINGERS 2 AND 3	BINDING SITE SIGNATURE	No. OF CLONES SELECTED AND SCREENED	BASES AT POSITIONS 4'X AND 5'X OF DNA SITES USED IN SELECTIONS	AMINO ACIDS SELECTED IN FINGERS 2 AND 3	BINDING SITE SIGNATURE	No. OF CLONES SELECTED AND SCREENED
4'X 5'X	F2 F3	4'X 5'X		4'X 5'X	F2 F3	4'X 5'X	
GG	6 -1123 R RLEY * R RSED R RHTH R RSSE R RSSA	G A T C G A T C	4 4 1 1 1	CC	6 -1123 T DPHN E HSKS E HRON S DRAN T DRAN	G A T C G A T C	1 4 1 1 2
AG	R QVTT * R QSGD R QLAT R QDAH R QRAS K QSTS	A G	4 5 2 1 1 1	AC	* Y RSQD V QVGH T QLAT	A A	3 2 3
TG	R SSGD R SASA * R NSGD R LVQN K TGAS R TPSG R TQTA R TSAA	T G	1 1 1 1 1 1 1 2	TA	N TSQA N TSAS K LAQT * E RSRD E RLRD V HLAT V HLTT L VGHH	T A	4 2 1 3 4 1 1 1
CG	R DTSV R DAST R DASA R DTSS	C G	3 2 1 2	CA	Y HPAT N HPAN E HHSN L DSRA	C A	2 1 9 2
GC	A RNHD * N RSTD * S RSTD S SRHS T RNST T RTST T TRY5 T RAQN	G C	2 3 6 1 1 1 3 1	GT	* T KSSD * K SSSD * K RSHD A RSSY T RSSS T RKTD	G T	2 1 1 1 1 1
AC	A QAAT A QGTN * V TSRD S QRGA S QSTT A TSSS T SSST	A C	4 1 4 1 2 1 1	AT	S QIST S QIGA T QYST T QSAS T QSQH T QTSH T QPGH T QDTH K QOST	A T	3 1 1 1 1 1 1 2
TC	D TISN N TSTA V TSSL V TSSI V TSNS * R GSND A TTSS T TAGS S TTSS S TSSA S LSTT G LSST	T C	1 1 4 1 3 1 1 1 1 1 1 1	TT	T TAST T TASH T TSSV T TSSA S HHTS S HAQT S HATT	T T	1 1 1 1 1 1 4
				CT	K OHSS T HPST S OSSR	C T	5 1 1

RELATIVE SIGNATURE STRENGTH ■ ≤1.00 ■ 0.80 ■ 0.60 ■ 0.40 ■ 0.20 □ 0.00

FIG. 3  
SUBSTITUTE SHEET (RULE 26)

4/6

**a**

AMINO ACID SELECTED AT POSITION -1

BASE PRESENT AT POSITION <sup>4</sup> X IN BINDING SIGNATURES	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	0	0	0	0	0	0	0	0	1	0	0	0	0	0	16	2	2	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0	21	1	1	1	0	0	0
T	0	0	0	0	0	1	7	0	0	4	0	1	0	2	5	3	21	1	0	0
C	0	0	10	0	0	0	7	0	0	2	0	0	0	0	2	0	0	0	0	0

↓ C                      ↓ T/C                      ↓ T                      ↓ A G(T)                      ↓ T

RECOGNITION PATTERNS

**b**

AMINO ACID SELECTED AT POSITION 6

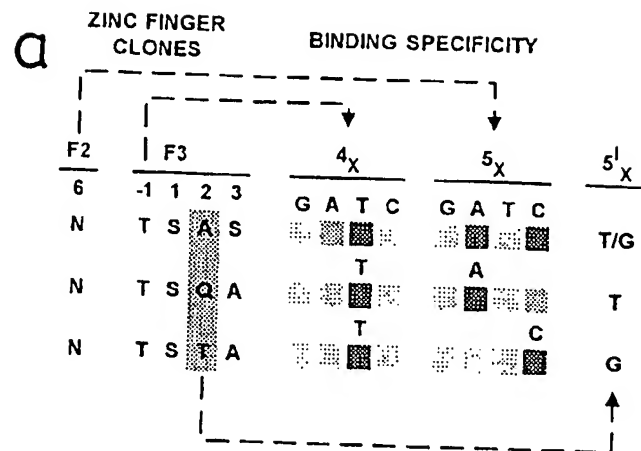
BASE PRESENT AT POSITION <sup>5</sup> X IN BINDING SIGNATURES	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
G	0	0	0	0	0	0	0	0	4	1	0	0	0	0	21	0	0	0	0	0
A	0	0	0	9	0	0	0	0	1	2	0	9	0	0	0	0	2	3	0	2
T	1	0	0	0	0	1	0	0	5	0	0	0	0	0	0	10	17	3	0	0
C	5	0	1	3	0	1	0	0	0	0	0	3	0	0	2	12	14	5	0	0

↓ C                      ↓ A/C                      ↓ G/T                      ↓ A/C                      ↓ G                      ↓ T/C C(A/T)

RECOGNITION PATTERNS

FIG. 4

5/6



**b**

AMINO ACID SELECTED AT POSITION 2      ZINC FINGER CLONES      CROSS-STRAND BINDING SPECIFICITY      No. OF CLONES SELECTED AND SCREENED

	F2	F3				5' <sub>X</sub>				
	6	-1	1	2	3	G A T C				
E	R	R	L	E	Y	[Pattern]				4
	R	R	S	E	D	[Pattern]				4
N	V	T	S	N	S	[Pattern]				3
	R	G	S	N	D	[Pattern]				1
Q	Y	R	S	Q	D	[Pattern]				3
	N	T	S	Q	A	[Pattern]				4
	K	L	A	Q	T	[Pattern]				1
	T	R	A	Q	N	[Pattern]				1
	E	H	R	Q	N	[Pattern]				1
	S	H	A	Q	T	[Pattern]				1
R	L	D	S	R	A	[Pattern]				2
	E	R	S	R	D	[Pattern]				3
	E	R	L	R	D	[Pattern]				4
	V	T	S	R	D	[Pattern]				4
H	A	R	N	H	D	[Pattern]				2
	S	S	R	H	S	[Pattern]				1
	T	D	P	H	N	[Pattern]				1
K	E	H	S	K	S	[Pattern]				4

FIG. 5

6 / 6

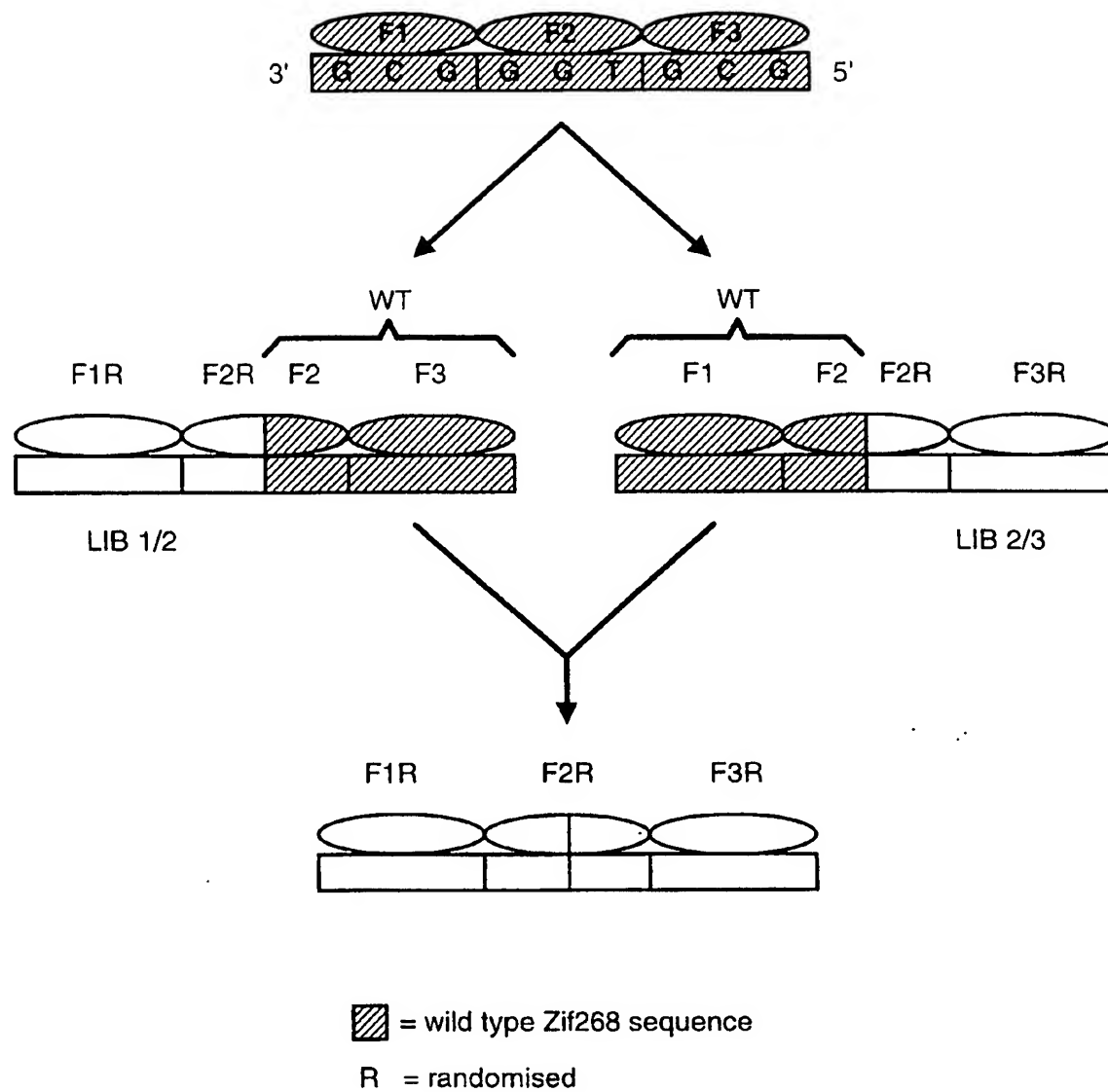


FIG.6

# INTERNATIONAL SEARCH REPORT

International Application No

PCT/GB 98/01510

## A. CLASSIFICATION OF SUBJECT MATTER

IPC 6 C12N15/10 C12N15/12 C07K1/04 C12N15/62 C12Q1/68  
C07K14/47 A61K48/00

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 C12N C12Q C07K A61K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, X	M. ISALAN ET AL: "Synergy between adjacent zinc fingers in sequence-specific DNA recognition" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, vol. 94, 27 May 1997, pages 5617-5621, XP002075337 WASHINGTON US see the whole document ---	1-25
X	CHOO Y ET AL: "DESIGNING DNA-BINDING PROTEINS ON THE SURFACE OF FILAMENTOUS PHAGE" CURRENT OPINION IN BIOTECHNOLOGY, August 1995, pages 431-436, XP000652427 cited in the application	1-6
Y	see the whole document see page 432, column 1 ---	10-25
-/--		

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

### \* Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- "&" document member of the same patent family

Date of the actual completion of the international search

1 September 1998

Date of mailing of the international search report

30/09/1998

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040. Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Cervigni, S

# INTERNATIONAL SEARCH REPORT

International Application No  
PCT/GB 98/01510

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT		
Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	<p>WO 96 06166 A (MEDICAL RES COUNCIL ;CHOO YEN (SG); KLUG AARON (GB); GARCIA ISIDRO) 29 February 1996 cited in the application see the whole document see page 5-6 see page 9, paragraph 3 see page 15, paragraph 2 see figures 2,7,15; tables 1,2 -----</p>	10-25

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/GB 98/01510

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9606166 A	29-02-1996	AU 3229195 A	14-03-1996
		CA 2196419 A	29-02-1996
		EP 0781331 A	02-07-1997
		JP 10504461 T	06-05-1998
-----			